Geometrical properties of connected search spaces for binary classification problem

Alexander Frey (sashafrey@gmail.com)
Konstantin Vorontsov (voron@forecsys.ru)

Computing Center RAS • Moscow Institute of Physics and Technology

25th European Conference on
Operational Research (EURO-XXV)
Vilnius, Lithuania • July 8 – 11, 2012

Empirical Risk Minimization

$(S, \mathcal{A}, \mathcal{P})$ a probability space;
$\mathscr{F}$ a class of measurable functions $f : S \to [0, 1]$ (losses)
Example: $S = X \times Y$, $f(x, y) = (g(x) - y)^2$.
Risk minimization

$$Pf \equiv \int_S fdP = \mathbb{E}f(x) \to min, f \in \mathcal{F}$$

Empirical risk minimization
$(x_1, \ldots, x_n)$ a sample of i.i.d random variables, $x_i \in S$

$$P_n f \equiv \frac{1}{n}\sum_{i=1}^{n} f(x_i) \to min, f \in \mathcal{F} \tag{1}$$

Empirical risk minimizer $\hat{f}$ — solution of (1)
Excess risk: $\varepsilon(\hat{f}) \equiv P\hat{f} - \inf_{f \in \mathcal{F}} Pf$.

Model Selection Problem:
Given a family $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}$ of nested function classes and sequence $\hat{f}_{n,k}$ of empirical risk minimizers on each class, select $\hat{f} = \hat{f}_{n,k} \in \mathcal{F}_k \subset \mathcal{F}$ with a "nearly optimal" excess risk
Approaches:

- **Penalization and oracle inequalities**, based on distribution dependent and data dependent bounds on $\varepsilon(\hat{f}_n)$ that take into account the "geometry" of $\mathcal{F}$, or
- in practice — **cross-validation**.

Empirical risk minimizer and mean overfitting

$\mathbb{X}^L = \{x_1, \dots, x_L\}$ a finite set of objects,
R — set of classifiers,
$\mathbb{X}^L = X^\ell \sqcup X^k$ decomposition of X_L into train and test sample,
$\nu(r, X^\ell) = \frac{1}{\ell} \sum\limits_{x_i \in X^\ell} I(r(x_i), y_i)$ — error rate of r on train sample,
$\hat{r} = \mu X^\ell = \operatorname*{argmin}\limits_{r \in R} \nu(r, X^\ell)$ — empirical risk minimizer,
$\delta(X^\ell) = \nu(\hat{r}, X^k) - \nu(\hat{r}, X^\ell)$ — *overfitting*,
With respect to decomposition $\mathbb{X}^L = X^\ell \sqcup X^k$, overfitting δ is a random variable,
$P_n \equiv \frac{1}{C_L^\ell} \sum\limits_{X^\ell}$ — empirical probability measure,
Cumulative distribution function: $Q(\varepsilon) = P_n\{\delta(X^\ell) \geqslant \varepsilon\}$,
Mean overfitting: $\bar{\delta} = P_n \delta(X^\ell) = \frac{1}{C_L^\ell} \sum\limits_{X^\ell} \delta(X^\ell)$

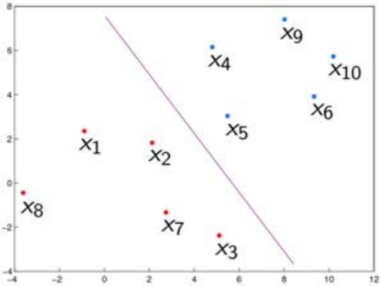
Decision with incomplete information

- Rows $\{x_1 \ \dots \ x_\ell, x_{\ell+1}, x_L\}$ — objects,
- Columns $\{r_1 \ \dots \ r_D\}$ — error vectors of classifiers.

	r_1	r_2	$\dots$	r_d	$\dots$	r_D
x_1	0	1	$\dots$	**0**	$\dots$	1
$\dots$	1	1	$\dots$	**0**	$\dots$	0
x_ℓ	0	0	$\dots$	**0**	$\dots$	0
$x_{\ell+1}$	1	1	$\dots$	**1**	$\dots$	1
$\dots$	1	0	$\dots$	**1**	$\dots$	0
x_L	0	0	$\dots$	**1**	$\dots$	0

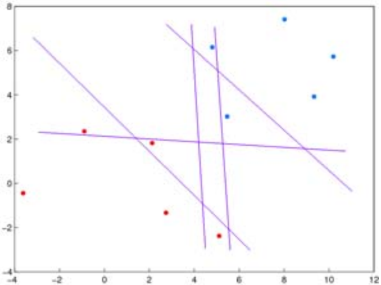
- $\{x_1, x_2, x_3\}$ — train sample,
- $\{x_4, x_5, x_6\}$ — test sample.

Example. Binary error matrix for a set of linear classifiers



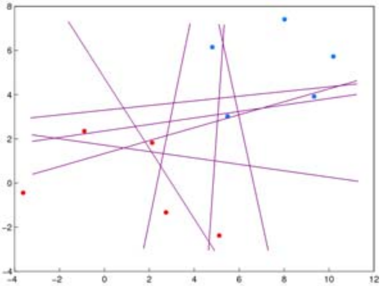
	layer 0
x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0

Example. Binary error matrix for a set of linear classifiers



	layer 0	layer 1				
x_1	0	1	0	0	0	0
x_2	0	0	1	0	0	0
x_3	0	0	0	1	0	0
x_4	0	0	0	0	1	0
x_5	0	0	0	0	0	1
x_6	0	0	0	0	0	0
x_7	0	0	0	0	0	0
x_8	0	0	0	0	0	0
x_9	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0

Example. Binary error matrix for a set of linear classifiers



	layer 0	layer 1					layer 2								
x_1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	...
x_2	0	0	1	0	0	0	1	1	0	0	0	0	0	0	...
x_3	0	0	0	1	0	0	0	1	1	0	0	0	0	1	...
x_4	0	0	0	0	1	0	0	0	1	1	0	0	0	0	...
x_5	0	0	0	0	0	1	0	0	0	1	1	1	0	0	...
x_6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	...
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	...
x_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...

Single classifier

Let R consists of single classifier.

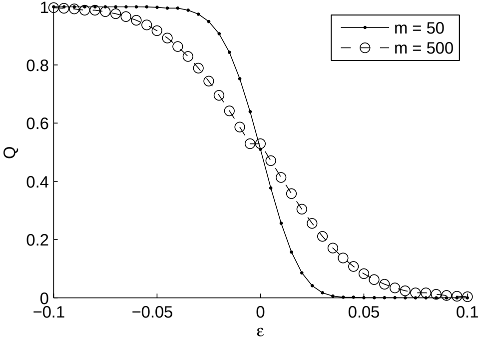


Figure: Cumulative distribution function $Q(\varepsilon) = P\{\delta(X^\ell) \geqslant \varepsilon\}$ of overfitting. $L = 1000$, $\ell = 250$.

Pair of classifiers

Let R consists of the pair of classifier.

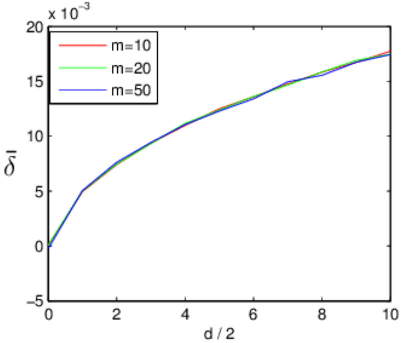


Figure: Mean overfitting $\bar{\delta}$ depending on the Hamming distance $d(r_1, r_2)$ in the pair of classifiers.

The maximal connected set of given diameter

Maximal set of classifiers with limited hamming diameter (2ρ) and fixed number of errors (m):

$$B_r^m(r_0) = \{r \in R\colon\ n(r, \mathbb{X}^L) = m,\ \text{and}\ \rho(r, r_0) \leqslant \rho\}.$$

R_n^m — set of n classifiers with m random errors.

r	$\lvert B_r^m \rvert$	$\lvert R_n^m \rvert$	δ
2	401	2	0.079
4	35.501	7	0.160
6	1.221.101	39	0.240
8	20.413.001	378	0.319

Table: Comparison of $\lvert R_n^m \rvert$ and $\lvert B_r^m \rvert$ that gives the sample $\bar{\delta}$. $L = 50$, $\ell = 25$, $m = 10$

Splitting and connectivity

Classical approach:

- δ-minimal sets:

$$\mathcal{F}(\delta) \equiv \{f \in \mathcal{F} \colon \varepsilon(f) \leqslant \delta\}$$

- L_2-diameter

$$D(\delta) \equiv \sup_{f,g \in \mathcal{F}(\delta)} (P(f-g)^2)^{1/2}$$

Combinatorial approach:

- Algorithms with low error rate on X_L

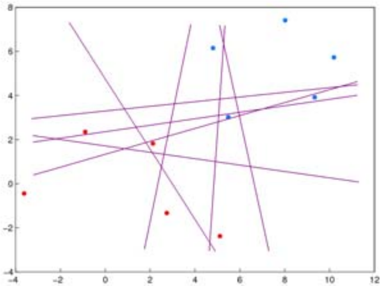
$$R(m) \equiv \{r \in R \colon n(r, X_L) \leqslant m\}$$

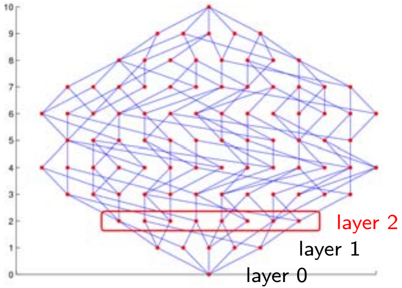
- Hamming diameter

$$D(m) \equiv \sup_{f,g \in R(m)} \rho(f,g)$$

$(\rho(r_1, r_2) = \sum_{x_i} [I(r_1, x_i) \neq I(r_2, x_i)])$

Error matrix and SC-graph for a set of linear classifiers





	layer 0	layer 1					layer 2								
x_1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	...
x_2	0	0	1	0	0	0	1	1	0	0	0	0	0	0	...
x_3	0	0	0	1	0	0	0	1	1	0	0	0	0	1	...
x_4	0	0	0	0	1	0	0	0	1	1	0	0	0	0	...
x_5	0	0	0	0	0	1	0	0	0	1	1	1	0	0	...
x_6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	...
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	...
x_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...

Splitting and connectivity profiles

Lets fix binary error matrix R.

- $\Delta_m = |\{r \in R \colon n(r, \mathbb{X}^L) = m\}|$ - splitting profile of R,
- $q(r_0) = |\{r \in R \colon \rho(r, r_0) = 1\}|$ - connectivity of classifier r_0,
- $\Delta_q = |\{r \in R \colon q(r) = q\}|$ - connectivity profile of R,
- $\Delta_{m,q} = |\{r \in R \colon q(r) = q \text{ and } n(r, \mathbb{X}^L) = m\}|$ - SC-profile.

SC-profile for linear classifiers

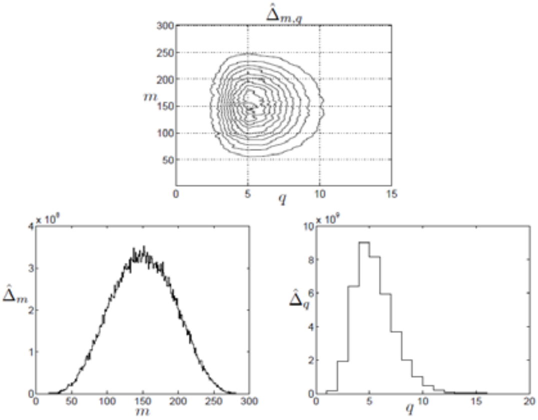
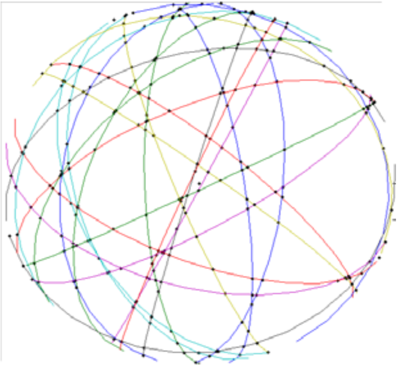


Figure: SC-profile for the set of linear classifiers in $\mathbb{R}^p$. $p = 5$, $L = 300$, $|R| = 2 \cdot 10^5$.

SC-profile for linear classifiers

- In R^3 consider the set $\mathbb{S}^2$ of linear classifiers $\{y = [\langle w, x\rangle \leqslant 0] \colon w \in \mathbb{S}^2, ||w|| = 1\}$.
- For a given object $x_0 \in R^3$ consider circle $\mathbb{S}^1 = \{w \in \mathbb{S}^2 \colon \langle w, x_0\rangle = 0\}$.

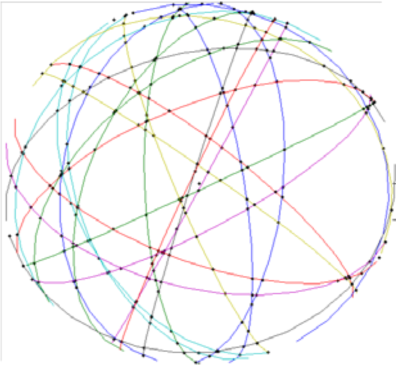


SC-profile for linear classifiers

This split $\mathbb{S}^2$ into cells.

- Each cell is the set of classifiers with identical error vectors,
- Edges between cells - classifiers that differs on one object.

Connectivity profile Δ_q doesn't depend on true classification!



- Binary classification problem: $Y = \{+1, -1\}$,
- $S_2 = \{e, h\}$ — group that acts on Y,
- S_2^L — group that acts on X_L (and hense on R)

Lemma

S_2^L doesn't change Hamming distance between classifiers:

$$\forall g \in S_2^L, \forall r_1, r_2 \in R \text{ holds } \rho(gr_1, gr_2) = \rho(r_1, r_2).$$

Theorem

The decomposition of SC-profile holds on average:

$$\frac{1}{2^L} \sum_{g \in S_2^L} \Delta_{m,q} = \Delta_q \times \frac{1}{2^L} \sum_{g \in S_2^L} \Delta_m$$

Conclusions

- Combinatorial approach deals with the same problems, as Statistical learning theory (model selection or sharp overfitting bounds),
- Instead of dealing with unknown underlying destribution, we study Complete Cross-Validation,
- We observe the same phenomena as in SLT — splitting and connectivity,
- We have proven that for binary classification problem connectivity is the geometrical propery of points, which doesn't depend on their target classes.